Quality Retrieval of the Empirical Literature: A Structured Approach

Leann L. Hankom Padmini Srinivasan

School of Library & Information Science
University of Iowa
Iowa City, Iowa 52242

The extent and growth of medical knowledge is tremendous. As of December 1990, the MEDLINE database had 6.5 million records in its collection, with approximately a third added during the last five years. Despite numerous innovations, retrieval technology has not kept pace with the exploding numbers of authors, articles, journals, books and conferences. A major limiting factor is that the basic mechanisms of retrieval systems almost uniformly rely on keyword representation and searching. These keywords are either assigned to texts perhaps with the assistance of special vocabularies, or appear naturally in these texts. In particular no structural or role information is preserved to connect these keywords to each other. This paper presents an alternative approach wherein role preserving, structured representations are used. This approach has the potential to increase retrieval quality.

INTRODUCTION

Experiments have always been a part of human endeavor. They allow investigators to experience and observe specific phenomena so as to increase our understanding of them. This empirical knowledge is then generally broadcast to a target audience through the established route of publications. Once retrieved, the results can be used in a wide variety of complex situations.

Given the rapid growth in medical knowledge, it is very challenging and difficult for practioners, researchers, educators, and students to keep abreast of empirical developments in their field. Today, researchers turn to online databases using keyword searches to obtain the information they need. However, in contrast with the simplistic keyword based representation and search mechanisms in standard text retrieval systems, questions that motivate health care practitioners to turn to such systems are very complex [7] as the following examples illustrate.

Example 1: What is the effect of dietary protein restriction on patients with diabetic nephropathy?

Example 2: Give me texts which describe the problems resulting in diabetes.

Example 3: Give me texts which describe problems caused by diabetes.

Example 4: Identify experiments which show that the temperature of an intravenous solution affects the level of pain felt by the patient.

These questions present highly specific as well as complex needs. In particular, they convey well defined roles for the key concepts. In example 1, 'dietary protein restriction' is the treatment and 'diabetic nephropathy' a central patient characteristic. When translated into the typical retrieval system's Boolean terminology: dietary protein restriction AND diabetic nephropathy, almost all of the role information is lost. The query will also retrieve texts in which diabetic nephropathy is a secondary patient characteristic or an observed variable, and where the dietary protein restriction is not the main treatment.

Unfortunately, Boolean retrieval cannot distinguish between examples 2 and 3. In example 4, Boolean retrieval cannot differentiate between studies which examine the relation between solution temperature and pain and those which actually find the effect that the user is interested in.

We contend that this situation regarding empirical knowledge could be improved if structured text representations, including role information replaced the bland keyword representations commonly used. In fact, we wish to show that at least in certain types of texts there is naturally occurring structure that can be utilized for more meaningful representation and retrieval. This is the motivation for our approach.

Our ideas have been implemented in a prototype

0195-4210/92/$5.00

retrieval system called Empiricist, implemented using hypertext. The text base in Empiricist consists of 157 empirical abstracts in two health domains: diabetes mellitus and intravascular therapy. Empirical abstracts are abstracts of articles reporting on experiments. Empirical knowledge has a critical role in the practice, teaching, and growth of health care [3]. The following examples illustrate the variety of situations for which our retrieval system may be found useful.

When designing an experiment, a health care researcher may need to identify similar experiments.

A researcher may need to examine studies which resulted in a particular conclusion.

A practitioner may need to choose between alternative treatments or tools for a given patient.

In essence, our approach gives the user the ability to think in terms of the desired experiment. Search specifications may exploit the various dimensions of empirical investigations. This feature gives Empiricist its advantage. We now describe our structured representation strategy and then the complex objects used to implement these structures.

STRUCTURED TEXT REPRESENTATION

Background

The representation used here is closely related to the naturally occurring structure of the text itself. Our approach derives from prior work, where we analyzed a group of empirical abstracts in the two health domains [4]. Structure here refers to the text's typical information components and their overall organization [3,8].

The previous study resulted in a hierarchical text structure or grammar for the texts analyzed which is presented in detail in [4]. Here we present only a brief overview as a background. Phrases within quotation marks indicate our text grammar components. Essentially, the typical empirical abstract presents the 'topic' and the 'design' of the underlying investigation. Topic may usually be expressed through the 'objectives' and the 'conclusions'. The experiment's 'design' may include one or more 'protocols', 'measures', and 'subjects'. Protocols are described by 'treatments' and 'procedures'. These abstracts also present important 'observations', 'data analyses' and 'conclusions' of the study. Components are composed of sub-components, thus yielding a hierarchical grammar. For example, 'subjects' may be described by their 'states', 'size' and 'sex'. In this way each text grammar component is broken down into its primitive information components. It should be noted that the empirical abstracts analyzed consistently focussed on these features.

Extracted Predicates

The first observation we made is that some parts of the text are not so useful for retrieval purposes. For instance, the 'observation' component consists of a 'measure' applied to one or more 'subjects' which yields a set of 'values'. For text retrieval, the fact that a measure was used is likely to be far more important than any other details about the observation. Therefore we use only select components from the text structure. In particular two types are selected to represent: (1) what the text is about and (2) the experimental design or methodology used in the investigation. These are extracted in the form of relational predicates whose structure closely match those of the corresponding text grammar components. These predicates are shown below in Figure 1.

Figure 1: Structure of Predicates

TOPIC predicates:

TOPIC(VARIABLE1, VARIABLE2, RELATIONSHIP, CONDITION, SOURCE)

DESIGN predicates:

TREATMENT(TYPE, SUBSTANCE, PURPOSE, TOOL, FREQUENCY, SITE, DURATION, DOSE)
PROCEDURE(TYPE, TOOL, PURPOSE, SITE, FREQUENCY, DURATION)
SUBJECT(TYPE, STATE, TREATMENT, PROCEDURE, SIZE, AGE, SEX)
MEASURE(ASPECT, NAME)
OVERALL_DESIGN(TYPE)

The TOPIC predicate's format results from a repeated observation in our previous analysis consistent with prior research [1,2,5]. These empirical studies predominantly investigate the nature of the relationship between pairs of key concepts under certain conditions. Key concepts, either abstract, as in a state, or concrete, as in a tool or

substance, translate into the two variables of the predicate. Also, the relationship argument is a directional one. Finally, the source argument specifies whether the TOPIC predicate derives from a statement describing objectives or one presenting conclusions. The DESIGN predicates represent key features of the methodology adopted.

The Complex Object Representation

The final representation of an abstract is in the form of a complex object (Figure 2). Complex objects use other objects for their description.

Figure 2: Complex Object Representation

REPRESENTATION
- TOPIC: TOPIC PREDICATE
  - VARIABLE 1 (V1)
  - VARIABLE 2 (V2)
  - CONDITION (C)
  - SOURCE (S)
- DESIGN:
  - MEASURES
    - MEASURE PREDICATE
      - NAME
      - ASPECT
  - TREATMENTS
    - TREATMENT PREDICATE
      - TYPE
      - SUBSTANCE (SU)
      - PURPOSE (PU)
      - TOOL (TL)
      - FREQUENCY (FR)
      - SITE (SI)
      - DURATION (DU)
      - DOSAGE (DO)
  - PROCEDURES
    - PROCEDURE PREDICATE
      - TYPE (TY)
      - PURPOSE (PU)
      - TOOL (TL)
      - FREQUENCY (FR)
      - DURATION (DU)
  - SUBJECT GROUPS
    - SUBJECT GROUP PREDICATE
      - TYPE (TY)
      - STATE (ST)
      - TR-TY (TR-TY)
      - PR-TY (PR-TY)
      - SIZE (SZ)
      - AGE (AG)
      - SEX (SX)
  - OVERALL DESIGN
    - OVERALL DESIGN PREDICATE
      - TYPE (TY)

Retrieval of complex objects might involve the retrieval of some or all sub-objects. Other researchers have used complex object representations for information retrieval. However, the focus is on the representation of texts by parts such as chapters, sections, sub-sections etc.

The root REPRESENTATION object is composed of a TOPIC object and a DESIGN object which are in turn broken down into smaller components. Predicate arguments are instantiated by one or more concepts, the most primitive objects in the composition hierarchy. Two types of objects are in the composition hierarchy: (1) objects whose immediate descendent objects are of different types (ex: DESIGN, PROTOCOL and TREATMENT PREDICATE) and (2) objects whose immediate descendants are of a single type (ex: set of PROTOCOLS, TOPIC, set of TREATMENTS). These objects are treated differently during retrieval. Figure 3 shows the complex object representation for one sample abstract.

Figure 3: Sample Abstract

Title: Administration of aspirin-dipyridamole reduces proteinuria in diabetic nephropathy.

Abstract: We assessed in a pilot study the effect on some aspects of renal function of 6 weeks' administration of a combination of aspirin-dipyridamole (990 mg/225 mg daily) administered on a double-blind crossover schedule in 16 insulin-dependent diabetic patients with nephropathy. Total 24-h urinary protein excretion (16 patients) was significantly reduced during aspirin-dipyridamole administration from a geometric mean (range) of 1.9 (0.4-7.7) g/24h to 1.4 (0.5-9.9) g/24h (2P less than 0.05). Indium-labelled platelet survival (eight patients), glomerular filtration rate and renal blood flow (eight patients) showed no significant change following aspirin-dipyridamole therapy, even though plasma creatinine concentration increased from 118 (65-371) to 130 (76-438) mumol/l (2P less than 0.05). Diabetic control and blood pressure remained unchanged throughout the study. Although the results showed that this treatment significantly reduced proteinuria in patients with diabetic nephropathy, the mechanism of action was

not entirely clear.

REPRESENTATION

TOPIC

TP1(V1: administration of aspirin-dipyridamole V2: proteinuria R: reduces C: diabetic nephropathy SOURCE: Conclusion)
TP2(V1: 6 weeks' administration of a combination of aspirin-dipyridamole V2: renal function C: insulin-dependent diabetic patients with nephropathy SOURCE: Objective)
TP3(V1: aspirin-dipyridamole V2: proteinuria R: reduced, C: diabetic nephropathy SOURCE: Conclusion)

DESIGN
MEASURES
MS1(NAME: 24-h urinary protein excretion)
MS2(NAME: Indium-labelled platelet survival)
MS3(ASPECT: rate of NAME: glomerular filtration)
MS4(NAME: renal blood flow)
MS5(ASPECT: concentration of NAME: plasma creatinine)
MS6(NAME: diabetic control)
MS7(NAME: blood pressure)
PROTOCOLS
PROTOCOL
TR1(SUBSTANCE:aspirin-dipyridamole DURATION: 6 weeks)
SUBJECT GROUPS
SB(STATE: diabetic nephropathy)
SB(STATE: insulin-dependent diabetes)
SB(SIZE: 16)
OD(TYPE: double-blind crossover)
OD(TYPE: pilot study)

The complex object representations from different texts are combined to form a conceptual network against which the retrieval strategies operate. Different complex objects may be connected since a given concept may be attached to a number of complex objects. Due to space limitations we do not describe this network any further. For this we refer the reader to [4] which describes Empiricist, a prototype implementation of these ideas.

Empiricist is implemented in HyperCard version 2.1 on a Macintosh IIci. Two different retrieval modes are built into it. Mode 1 allows the user to specify the query by a set of concepts. In mode 2, the user has the option to fill in predicate templates, i.e. constrain concepts to specific roles. This second option is valuable for a searcher who wants to describe the relevant texts via the different dimensions of an empirical investigation.

CONCLUSIONS

The structured representation approach proposed here is better suited for some of the complex queries faced by users. These complex object representations derive from the naturally occurring structure in texts. Most importantly our approach allows the user to think in terms of features of the desired experiment. In this way it provides a more sophisticated approach to retrieval for users' complex queries.

We conclude with a word about the extraction of predicates. In our current work this extraction is done by project participants. We suggest that given the intuitive nature of these predicates, it should similarly be possible for authors or indexers to extract them from abstracts. In fact our current goal is to conduct an experiment which will test this suggestion. We are also in the process of evaluating Empiricist.

Reference

[1] Graves, Judith Rae. A Research-Knowledge System (ARKS) for Storing, Managing, and Modeling Knowledge from the Scientific Literature. *Advances in Nursing Science* 13, 2 (1990), pp. 34-45.

[2] Horowitz, R. S. and Weiner, J. M. Article Processing After Retrieval. *MedInfo*. 1986. Eds. R. Salamaon, B. Blum, and M. Jorgenson. Elsevier Science Publishers, North-Holland.

[3] Kintsch, W. and van Dijk, T. A. Toward a Model of Text Comprehension and Production. *Psychological Review* 85 (1978), pp. 363-394.

[4] Rama, D. V. and Srinivasan, Padmini. An Investigation of Content Representation using Text Grammars. To appear in *ACM Transactions on Information Systems* (1992).

[5] Rennels, Glenn D. *A Computational Model of Reasoning from the Clinical Literature*. Lecture Notes in Medical Informatics. 1987. Eds. P. L.

Reichertz and D. A. B. Lindberg. Springer-Verlag.

[6] Scura, Georgia and Frank Davidoff. Case-Related Use of the Medical Literature: Clinical Librarian Services for Improving Patient Care. *Journal of the American Medical Association* 245, 1 (1981), pp. 50-52.

[7] Shapiro, Alan R. Knowledge Retrieval and the Medical Information Sciences. In *Frontiers of Medical Information Sciences*. 1988. Ed. R. L. Kuhn. Praeger.